Tom Kalvelage, EDC DAAC

This white paper offers some thoughts on several of the topics.

First, overall, NASA is used in the paper as a monolithic agency, and it is not. So, throughout this white paper, I will refer to the "NASA Project". This entity is an implementing organization something like ESDIS that is responsible for the overall accomplishment of a mission.

Standards and Interface Processes for Future Missions Study

First, science communities are not homogeneous, so a science community will not be able to determine an interface. For example, one science community determined what the best projection was for MODIS products, but another science community did not think that was correct. The interface designer will have to decide what sort of community is the target (data producers, NASA scientists, commercial users, and so on), and get a representative sample from that community. A panel or organization can determine an interface, but a community cannot, unless there is an election by the entire community.

Second, a science organization can be considered to be truly 'in the science community' only if it is solely concerned with science. The MODIS land product format case was a terrific lesson learned for SEEDS. What was considered at the time to be the science community was in charge of MODIS production and could produce any format they wished. However, because this particular science community was heterogeneous (land, oceans, and atmospheres) and directly responsible for production as well as science, an interface/projection was designed that users did not like (at least in the case of land users).

So, the NASA Project in the future will need to decide if a particular science community is representative of a larger community that it wishes to reach. If a science organization/community is in charge of something besides science, it can't be considered to be 'the' science community because it has a conflict of interest that ordinary science users don't have. In that case, a science advisory panel should be set up to ensure that the science organization's non-science responsibilities aren't leading it astray from the community it represents.

Third, NASA has been quite emphatic recently that it is concerned with the amount of data being distributed, and wishes to get more data out. In that case, standards and interfaces should be determined by anyone who is a user - individuals, commercial companies, other government agencies, etc. So, interfaces can't be assigned to the science community as a default, but to a user community.

Fourth, standards are of two types, technical and operational. Technically, we have had the most success with existing standards. New standards don't always work and they always require special training (even for users, when we involve them). In NASA missions, the data itself is generally new, so using existing standards reduces the number of variables (generally a good thing in science) for the user and for the operations and maintenance (O&M) staff.

For the users, using new formats and projections means the data is not as accessible, and we expend more user services effort to support them - when they bother to call in instead of just refusing to use the data until it is more useable.

For O&M staff, we have had good luck using commercially available databases, protocols, and interfaces (e.g. ftp) wherever possible. Not only is it easier to hire staff with this knowledge, but it is easier to retain them, and so we generally have more experienced staff on hand - which leads to better RMA and more satisfied users. Using new standards (e.g. DCE) and interfaces (e.g. polling server) adds more variables and work for the O&M folks. It can be done, but the extra workload required means it should only be done where necessary (which it is sometimes, particularly with NASA missions).

Operational standards and interfaces are at least as important as technical standards and interfaces. Operationally, we have found a few standards that work very well if used, but add a lot of work, lost data, and friction if not mandated (because a lot of time is spent in negotiations and learning the same lessons over and over), or worse, not used at all.

> First of these is that the data producer must be (and really is - no one else can do it) responsible for getting their data to, <u>and received by</u>, the user. If the user (who may be an end user or an entity in the production chain, or the true end user) does not get the data, it is the producer's job to find that out and retransmit. If at all possible, this should be automated, particularly at high data rates.
>
> Second of these is that some one entity must be visibly in charge of the overall system during system operations. A large system-of-systems will not organize itself in any deterministic way. If throughput, RMA, response time, etc, is not a high priority, this is not a requirement, but in most production systems that is not the case. If a system component decides it is going to change it's operation or requirements, someone needs to evaluate those changes' effect on the end-to-end system.
>
> The third operational interface is that there must be a visible and agreed-upon source of actual build-to requirements, and preferably, a list of 'wish list' requirements. In a large multi-component system, individual components will start to diverge from the overall system, and confusion will increase as requirement sets multiply.
>
> The fourth operational interface is that there must be close communication between any adjacent components in the overall system, and, if necessary, the one entity in charge of overall operations (mentioned in the second ops interface above) may choose to participate in this. This helps correct problems in hours, rather than weeks or months

(or never). Again, if the ops requirements (RMA, timeliness, etc, etc) are not tight, this may not be necessary.

If there is time and funding, standards and interface processes typically work best if there is a focus on who is being served by the standard or interface, and what the requirements are (e.g. RMA, timeliness, data loss, etc). Stakeholders get some buy in, someone makes a decision, reaction is solicited, and then a final decision is made. If there is not time or funding, someone is designated the expert, they make the call and you live with it.

It generally doesn't work well if there is endless discussion and a decision is never made, or if the expert makes a call without considering all the requirements or users.

Life Cycle / Long Term Archive

I'll address the archive manager section.

The critical issue is that the data provider (that is, whoever on the data provider's side who has the funding and authority to make the transfer) and the LTA must have direct communication and negotiation with each other. The active archive cannot commit funding or establish policy for either the data provider or the LTA. The active archive can greatly assist the communication between the two entities, and help with all the technical and most of the policy issues. But the active archive, unless it has been given the funding and authority to make the transfer, cannot represent either one. Most particularly, the active archive cannot be expected to provide the initiative for making the LTA transfers execute. Without funding or authority, the active archive can plead, beg, cajole, argue, and/or implore, but can't actually force anything to happen.

It would be best if the data provider and the LTA enter into negotiations about content standards, data formats, and so on while the mission was being planned. The LTA cannot be expected to follow requirements unless funded to do so by either their own funding means or by the mission. By entering into discussions early, it will be easier to plan funding, and expectations at least can be managed.

It is not expected that at a very early stage that the LTA will know what it's funding or practice will be at the end of a proposed NASA mission (plus final reprocessing) - that could be ten years or more away. So, each data product will likely be handled on a case-by-case basis. However, by having the early discussions, the data providers and the LTA would encourage the development of common standards, develop their working relationships, and understand better what funding is required by who, and when.

The major cost drivers for migration from SIPS to active archive to LTA are essentially complexity, volume, level of service required (and the suitability of the data to support that level of service), and policy negotiation.

Complexity: complexity of product, whether software is required to produce them, whether the software requires hard-to-maintain commercial hardware or software, whether products require lengthy science training by the data producers for the LTA staff, whether interface and fault recovery is automated, whether volumes require a lengthy migration process (e.g. years), etc.

Volume: sheer number of terabytes, number of products, amount of processing power required for processing.

Level of Service: user services, browse, currency, turn-around time, user interface, expectation of user community.

Policy negotiation: how much effort is required to negotiate and agree to the other three factors, and then acquire funding.

I believe the major cost driver will change for each data set, as each data set is different. Costs can be contained by agreeing to what is required early, by making the products (and associated software) easier to maintain, by having common data requirements (where possible) common across missions, and by requiring less.

I have no miracle guidelines for data providers. In general, I would recommend to data providers that they produce data with clear QA information, useful file and granule names, simple associated browse, easy to understand formats and organization, and clear packaging (how a product is grouped). This will make data easier for the users to use.

These products should be transferred from organization to organization through an automated interface and process if the data set is large enough to warrant it. Knowledge transfer should be done as rigorously as possible (don't ask the LTA to sift through 100 shelf-feet of documents - they will need to be briefed).

The biggest near-term computer technology for LTAs is the decreased cost of on-line storage. The more data that is immediately and easily available and accessible, the more use the older data sets will get. However, since data producers are increasing their data rates to keep up with technology, I don't think there is a silver bullet computer technology to help the transfers.

It is possible that a technology advance in the soft sciences or a cultural/political change could occur would allow us to better manage the organizational interactions required of the data transfers. That would have a huge impact on the data transfers, but it is unlikely to happen.